Molecules as Documents of Evolutionary History

By Emile Zuckerkandl and Linus Pauling

Gates and Crellin Laboratories of Chemistry
California Institute of Technology
Contribution No. 3041

1. The chemical basis for a molecular phylogeny.

Of all natural systems, living matter is the one which, in the face of great transformations, preserves inscribed in its organization the largest amount of its own past history. Using Hegel's expression, we may say that there is no other system that is better "aufgehoben" (constantly abolished and simultaneously preserved). We may ask the questions where in the now living systems the greatest amount of their past history has survived and how it can be extracted.

At any level of integration, the amount of history preserved will be the greater, the greater the complexity of the elements at that level and the smaller the parts of the elements that have to be affected to bring about a significant change. Under favorable conditions of this kind, a recognition of many differences between two elements does not preclude the recognition of their similarity.

One may classify molecules that occur in living matter into three categories according to the degree to which the specific information contained in an organism is reflected in them :

(1) Semantophoretic molecules or semantides -- molecules that carry the information of the genes or a transcript thereof. The genes themselves are the primary semantides (linear "sense-carrying" units). Messenger-RNA molecules are secondary semantides. Polypeptides, at least most of them, are tertiary semantides.

(2) Episemantic molecules -- molecules that are synthesized under the control of tertiary semantides. All molecules built by enzymes

in the absence of a template are in this class. They are called episemantic because, although they do not express extensively the information contained in the semantides, they are a product of this information.

(3) Asemantic molecules -- molecules that are not produced by the organism and therefore do not express, either directly or indirectly (except by their presence, to the extent that this presence reveals a specific mechanism of absorption), any of the information that this organism contains. However the organism may often use them, and may often modify them anabolically and thus change them into episemantic molecules to the extent of this modification. The same molecular species may be episemantic in one organism and asemantic in another. Vitamins constitute examples. Simple molecules such as phosphate ion and oxygen also fall into this category. Macromolecules found in an organism for any length of time are never asemantic, viruses excepted. Viruses and other "episomes" (Wollman and Jacob, 1959) are asemantic when present in the host cell in the vegetative, autonomous state; they are semantophoretic when integrated into the genome of the host.

Products of catabolism are not included in this classification. During the enzymatic breakdown of molecules, information contained in enzymes is expressed, but instead of being manifested in both the reaction and the product, this information is manifested in the reaction only. Since we are considering products, catabolites as such are non-existent with respect to the proposed classification.

The relevance of molecules to evolutionary history decreases as one passes from semantides to asemantic molecules, although the latter

may represent quantitative or qualitative characteristics of groups. As such they are, however, unreliable and uninformative. It is plain that asemantic molecules are not worthy of consideration in inquiries about phylogenetic relationships.

Neither can episemantic molecules furnish the basis for a universal phylogeny, for such molecules, if universal, are not variable (ATP), and, if variable, are not universal (starches). It appears however possible a priori that parts of the phylogenetic tree could be defined in terms of episemantic molecules. An attempt in this direction has been made for instance on the basis of carotenoids in different groups of bacteria (cf. Goodwin, 1962). It is characteristic of such studies that they need independent confirmation. Such independent confirmation may be obtained by direct or indirect studies of semantides. In relation to a number of organic molecules, such as vitamin B_{12}, organisms as far apart on the evolutionary scale as bacteria, flagellates, and higher vertebrates differ, not in that the compound is present or absent, required or not required, but in the prevalent "pattern of specificity" (Hutner, 1955). By this is meant the measure of functional effectiveness of compounds closely similar to but not identical with the one that is actually present. Thereby the difference of organisms in relation to the organic molecules under consideration is reduced to differences in enzymes and, in the last analysis, to the difference in primary structure of polypeptide chains. Because of this relationship to studies of semantides, it is possible that the establishment of different patterns of specificity is one of the best uses to which episemantic and asemantic molecules may be put in phylogenetic studies.

Whereas semantides are of three types only (DNA, RNA, polypeptides), episemantic molecules are of a great variety of types. Their interest for phylogeny is proportional to their degree of complexity. Polysaccharides such as cellulose are large molecules, but their complexity is small because of the monotonous repeat of the same subunits. In fact, not only episemantic molecules, but also semantides vary in their degree of complexity. The complexity of semantides is largest in the case of large globular polypeptide chains and smallest in the case of structural proteins characterized by numerous repeats of simple sequences. There may be a region of overlap of semantides with the lowest degree of complexity and of episemantic molecules with the highest degree of complexity. The former, however, will still contain more information than the latter about the present and the past of the organism. Indeed, episemantic molecules are mostly polygenic characters, in that enzymes controlled by several distinct structural genes have to collaborate in their synthesis; moreover, they express the information contained in the active centers of enzymes only, and in no other enzymatic region; and even then express this information ambiguously; i.e., probably with considerable "degeneracy". There is thus a great loss of information as one passes from semantides to episemantic molecules. Incidentally, one cannot yet be sure that all polypeptides are semantides. Some, especially among the small ones, but also among the large structural ones, may be episemantic. Thus, it has not been possible to split glutenin into subunits of reproducible molecular weight (Taylor and Cluskey, 1962). This raises the suspicion that glutenin might not be produced as the transcript of a template.

Because tertiary semantides (enzymes) with different primary structures can lead to the synthesis of identical episemantic molecules as long as the active enzymatic sites are similar, wrong inferences about phylogenetic relationships may be drawn from the presence of identical or similar episemantic molecules in different organisms. Amylopectins in plant starches and animal glycogens are very similar, yet we may expect (a point it will be of interest to verify) that the amino-acid sequence of the enzymes responsible for the synthesis of these polysaccharides in animal and plant kingdoms is very different. Moreover a similar end-product, in the case of episemantic molecules, may be obtained by different pathways, so that not even the active sites of the enzymes involved need to be similar. The synthesis of nicotinic acid and that of tyrosine are carried out via different pathways in bacteria and in other organisms (cf. Cohen, 1963). Therefore, the presence of these molecules in no way points to a phylogenetic relationship between bacteria and these other organisms. The number of possible historical backgrounds to the presence of a molecule synthesized by an organism will tend toward unity only as the number of enzymes involved in the synthesis of this molecule increases significantly. It is not likely that a whole pyramid of enzymatic actions has been built more than once or twice during evolution. This consideration implies that the best phylogenetic characters among episemantic molecules are not just the most complex molecules but, among these, the ones that are built from the least complex asemantic molecules.

The preceding discussion suggests that the most rational, universal, and informative molecular phylogeny will be built on semantophoretic molecules alone. Evolution, in these molecules, seems

to proceed most frequently by the substitution of one single building stone out of, say 50 to 300 for polypeptides or, on the basis of a triplet code, 150 to 900 for the corresponding nucleic acids. Even these small changes can have profound consequences at higher levels of organic integration, through an alteration of the established pattern of molecular interaction. Therefore, in macromolecules of these types there is more history in the making and more history preserved than at any other single level of biological integration.

In previous communications (Zuckerkandl and Pauling, 1962; Pauling and Zuckerkandl, 1963) we have discussed ways of gaining information about evolutionary history through the comparison of homologous polypeptide chains. This information is threefold: (1) the approximate time of existence of a molecular ancestor common to the chains that are being compared; (2) the probable amino-acid sequence of this ancestral chain; and (3) the lines of descent along which given changes in amino-acid sequence occurred. The first type of information is obtained in part through an assessment of the overall differences between homologous polypeptide chains. The second and third types of information are obtained through a comparison of individual amino-acid residues as found at homologous molecular sites. Our purpose was to spell out principles of how to extract evolutionary history from molecules, rather than to write any part thereof in its final form -- an attempt that would require more information than is presently available even in the case of hemoglobins.

Beside the analysis of amino-acid sequence of a greater number of homologous polypeptide chains, two other sources of knowledge will help in retracing the evolutionary history of molecules. One is a consideration of the genetic code, to assess whether the passage of one

character of sequence to another could have occurred in one step, to discover possible intermediary states of sequence, and to evaluate to a better approximation than by the simple comparison of the amino-acid sequence of two homologous polypeptide chains the minimum number of mutational events that separate these two chains on the evolutionary scale.

A second is the study of three dimensional molecular models, to permit one to make predictions about the effects of particular substitutions and, on the basis of the transitions allowed by the genetic code, to exclude some substituants, as incompatible with the preservation of molecular function, from the list of possible evolutionary intermediates.

This cursory outline of methodology in chemical paleogenetics applies directly to the analysis of polypeptide chains only. Although techniques are not yet available for a thorough investigation of sequence in other types of semantides, it is of interest to examine the relationship between the different types of semantides with respect to the information they contain.

2. Cryptic genetic polymorphism through isosemantic substitution.

For any one corresponding set of molecules, the three scripts used by nature in semantophoretic molecules, the DNA-, RNA- and polypeptide scripts, represent largely, but presumably not exactly, the same message. Errors of transcription (Pauling, 1957) are presumably only a minor cause of this lack of congruence. In view of the "degeneracy" of the genetic code, many amino acids appearing to be coded for by more than one type of codon (Weisblum et al., 1962; Jones and Nirenberg, 1962), one must assume that information is lost in the passage from secondary (RNA)

to tertiary (polypeptide) semantides. Moreover many primary semantides may not be transcribed : there are significant stretches of DNA that are apparently not expressed in polypeptide products, and the base sequence along these stretches may represent important documents about the history of the organism as well as its present organization and potentialities.

The degeneracy of the genetic code, then, leads one to predict the existence of isosemantic heterozygosity, namely of differences in base sequence in allelic stretches of DNA that do not lead to differences in amino-acid sequence in the corresponding polypeptide chains. The base sequence of a codon may be changed, but the "sense" of the "word", in terms of amino acids, may remain the same. The same inference has been drawn independently by Richard T. Jones (personal communication and reference).

Eck (1963) proposes that one of the three letters of each codon, perhaps the middle letter, is recognized by transfer-RNA's only as "purine" or "pyrimidine"; that is, according to the bulk of the molecule rather than to its exact species. Thus shifts between adenine and guanine or between cytosin and uracil in the middle letter of messenger-RNA codons will not be heeded by transfer-RNA. If this is so, one must distinguish two levels of crypticity. Some base substitutions will remain cryptic, unexpressed at the level of the polypeptide chains, but will be recognized at the level of transfer-RNA (secondary crypticity). Other base substitutions will remain cryptic at the level of both the polypeptide chain and the transfer-RNA (primary crypticity). (A third, more superficial level of crypticity was often referred to in the past, namely cryptic amino-acid substitutions in polypeptides, substitutions that were supposed to actually exist, but not to be detected by available chemical

means.) According to Eck's code, primary crypticity should exist for every amino acid, because of the peculiar role tentatively attributed to the middle letter of the codon, and secondary crypticity should exist for eleven amino acids out of twenty. Amino acids that occur with high frequencies usually seem to have degenerate codes. The opportunities for isosemantic substitution and cryptic genetic polymorphism, even of the secondary type of crypticity, should therefore be very widespread indeed.

As is well known, the abnormal human hemoglobins HbS and HbC differ from HbA in that a valyl residue replaces a glutamyl residue in the β-chain of HbS at the sixth position from the amino-end, whereas a lysyl residue replaces the same glutamyl residue in HbC (Ingram, 1957; Hunt and Ingram, 1958, 1959). According to the proposals for the genetic code made by Jukes (1962), by Wahba et al. (1963) and by Eck (1963), the shift from valine to lysine is one of the rare ones that require three base pair substitutions in DNA and therefore, presumably, three mutational steps. If correct, this conclusion would render unlikely the hypothesis, previously formulated by one of us (Pauling, 1961), that HbC is derived from HbS rather than from HbA. The three genetic codes, on the other hand, are compatible with a one-step transition between HbA and HbS as well as between HbA and HbC. According to Eck's code - not according to the other proposals for a genetic code mentioned above - the valine of HbS and the lysine of HbC must however have derived from two distinct isosemantic codons for glutamic acid in HbA (Fig. 1). Whether or not Eck's code will, in the end, be shown to be correct in this respect, we may accept it provisionally for the sake of this discussion. Indeed, even if HbS and HbC are not the products of mutations in two isosemantic

codons, other cases of this type are likely to be found in the future.

If the situation is as represented in Fig. 1, the two isosemantic codons for glutamic acid (which are actually, according to Eck, resolvable into four isosemantic triplets, AAG, AGG, UAG and UGG) must be thought to have at one time been widespread in the human population, and may even today constitute a case of perhaps widely occurring cryptic genetic polymorphism. A search for it might be made in particular among individuals who appear to possess different isoalleles of HbA, in the nomenclature of Itano (1957), namely different heritable relative levels of HbA production. Because all transfer-RNA's that would correspond to all possibilities of a degenerate code might conceivably not be available in excess in certain organisms or in certain tissues of an organism, isosemantic substitutions may lead to increased or decreased rates of polypeptide synthesis. Thus there could exist an operator-independent change in rate of protein synthesis through base-pair substitutions in structural genes. Pauling ( ) and Itano (1957) used to think that the existence of isoalleles is probably linked to cryptic substitutions. In terms of cryptic amino-acid substitutions, this hypothesis has ceased to be as likely as it appeared to be, at least as a very generally applicable explanation. Yet it may be reintroduced on the basis of cryptic base substitutions in DNA and RNA.

If the scarceness of some species of isosemantic transfer-RNA's, can affect the rate of synthesis of polypeptides, the synthesis of a single polypeptide chain should not proceed at constant speed along the chain, but so to speak in jerks, with sudden decelerations at molecular sites where the codons happen not to correspond to species of transfer-RNA that are present in excess. No evidence for or against this effect is available to our knowledge.

One may ask the question whether cryptic isosemantic substitutions may offer a possible alternate explanation, beside those already proposed, of the thalassemic inhibition of certain hemoglobin chain genes. A single isosemantic substitution may form a bottle-neck in synthesis, but an additive effect of such substitutions is also possible. The human δ-chain may be one that is universally "thalassemic" in this sense. This interpretation of thalassemia or of the low rate of synthesis of δ-chains would imply that normal amounts of the corresponding messenger-RNA's are produced and occupy a significant percentage of the available ribosomal sites without leading to much synthesis. Such an effect, is not very likely, especially in thalassemia. Indeed, an extensive survey of the literature has shown in heterozygotes for α-thalassemia or β-thalassemia a "compensatory" increase of mean absolute amounts per cell of the chain synthesized under the control of the allelic gene, whether this allelic gene be normal or structurally abnormal (unpublished). This observation suggests that more ribosomal sites have become available to messenger-RNA produced by a single allele than is the case when both alleles are normally active. If correct, this interpretation would imply that the output of messenger-RNA by thalassemic hemoglobin genes is actually reduced, and that the block of polypeptide synthesis in thalassemia is at the genic level rather than at the level that involves the action of transfer-RNA. On the other hand one may surmise that the low rate of synthesis of δ-chains is correlated with a low genic output of messenger-RNA more probably than with low synthetic efficiency at the ribosomal level. Moreover, the hypothetical effect on rate of synthesis of isosemantic substitutions is not supported by Eck's code for hemoglobin S, if it is assumed that HbS has arisen from HbA. According to Eck, there

is indeed only one codon for valine recognizable by transfer-RNA, namely UYG (i.e., UCG and UUG; Y stands for "pyrimidine"). One can therefore not say, without resorting to an auxiliary hypothesis, that the apparent slower relative rate of HbS synthesis as compared to HbA synthesis in HbA/HbS heterozygotes is perhaps due to the appearance of a codon whose corresponding transfer-RNA is present in limiting amounts. The auxiliary hypothesis is that a given kind of transfer-RNA is not entirely indifferent to the exact chemical species of the central purine or pyrimidine in a codon. It is possible that the exact chemical species of the presumed middle "letter", while without action on the coding, influences the rate of synthesis of the polypeptide. However, very recent evidence (Levere and Lichtman, 1963) suggests that the rate of synthesis of HbS may in reality not be inferior to that of HbA. The present status of these problems is uncertainty.

Although there is therefore no evidence in favor of considering isosemantic substitutions as a significant factor in the regulation of the rate of polypeptide synthesis, the possibility is not ruled out and should be kept in mind as furnishing a basis for Itano's idea, expressed here in a slightly modified way, that rate of synthesis and structure, at the level of a given structural gene, are intimately linked (Itano, 1957).

If isosemantic substitutions recognized by transfer-RNA actually exerted an effect on rate of polypeptide synthesis, one would expect natural selection to act quite strongly on such substitutions. If natural selection did not act on the other postulated type of isosemantic substitutions, those of "primary crypticity", not recognized by transfer RNA, the occurrence of such substitutions would be random. It would be more probable that some effect is present and that natural selection acts here

also. A possible effect of the exact chemical species of the presumed middle letter of a codon on rate of synthesis has just been mentioned. Other effects might be considered. In particular, it has been shown that the frequency of crossing over is inversely related to the degree of heterozygosity in a chromosome pair (Stadler and Towe, 1962). Via this mechanism isosemantic substitutions of primary crypticity as well as those of secondary crypticity may have far reaching effects on population genetics.

One may also examine the possibility that isosemantic substitutions have some effect on evolutionary stability. Benzer (1961) has pointed out that, since AT base pairs are held together much less strongly than GC base pairs, a genetic region rich in AT pairs will tend to be more subject to substitution. By selecting for isosemantic triplets rich in GC and low in AT content, an organism might reduce its mutation rate without changing the structure of any of its proteins. Whether such an effect would be significant enough to influence the rate of evolution remains to be seen. It will be of interest to compare the base composition of DNA from "living fossils" such as Lingula or Limulus with base composition in more rapidly evolving animals.

Finally, isosemantic substitutions in those regions of DNA that carry out the function of operators (Jacob and Monod, 1961) might well lead to a modification of the stereochemical relationship between the operators and the repressor molecules. Thereby such isosemantic substitutions would have an effect on rate of polypeptide synthesis, distinct from the operator-independent effect discussed earlier.

Due to isosemantic substitutions, there probably is more evolutionary

history inscribed in the base sequence of nucleic acids than in the amino-acid sequence of corresponding polypeptide chains. By its implications, a degenerate code thus emphasizes the role of nucleic acids as "master molecules" over polypeptides, (a role still doubted by some (Commoner, 1962)), even though polypeptides may interact with nucleic acids to regulate the rate of synthesis of both polypeptides and nucleic acids. All the potentialities of an individual may be assumed to be inscribed in polypeptide chains that are actually synthesized, or could be synthesized, by the cells under certain circumstances, and in the structures that control the actual and potential rates of this synthesis. Yet it appears conceivable, since equal rates of synthesis under the control of distinct but isosemantic codons are possible, that the individual contains information, not only, as we know, beyond that which it actually uses for its realization, but even beyond that which defines its potentialities. This part of its "being", necessarily cryptic in terms of the phenotype, would at best be expressed only in relation to the evolution of the species.

REFERENCES

Benzer, S., Proc. Natl. Acad. Sci. 47, 403-415, 1961.

Cohen, S. S., Science 139, 1017-1026, 1963.

Commoner, B., in : "Horizons in Biochemistry", M. Kasha and B. Pullman, eds., Academic Press, New York, 1962, pp. 319-334.

Eck, R. V., Science 140, 477-481, 1963.

Goodwin, T. W., in : "Comparative Biochemistry", M. Florkin and H. S. Mason, eds., Vol IV, Academic Press, 1962, pp. 643-675.

Hunt, J. A. and Ingram, V. M., Nature 181, 1062-1063, 1958.

Hunt, J. A. and Ingram, V. M., Nature 184, 870-872, 1959.

Hutner, S. H., in: "Protozoa", S. H. Hutner and A. Lwoff, eds., Vol. 11, Academic Press, New York, 1955,

Ingram, V. A., Nature 189, 704-708, 1961.

Itano, H. A., Advances in Protein Chem. 12, 215-268, 1957.

Jacob, F. and Monod, J., Cold Spring Harbor Symposia Quant. Biol. 26, 193-211, 1961.

Jones, O. W. and Nirenberg, N. W., Proc. Natl. Acad. Sci. 48, 2115-2123, 1961.

Jones, R. T., in : "Symposium on Foods : Proteins and their Reactions", H. W. Schultz and A. F. Anglemier, eds., The Avi Publishing Co., in press.

Jukes, T. H., Proc. Natl. Acad. Sci. 48, 1809-1815, 1962.

Levere, R. D. and Lichtman, H. C., Blood 22, 334-341, 1963.

Pauling, L., in : "Arbeiten aus dem Gebiet der Naturstoffchemie. Festschrift Arthur Stoll", Birkhäuser, Basel 1957, pp. 597-602.

Pauling, L., Rudolf Virchow Memorial Lecture, New York, 1961; Proceedings of the Rudolf Virchow Medical Society, 21, 1963 (S. Karger AG, Basel).

Pauling, L. and Zuckerkandl, E., Acta Chem. Scand., in press.

Stadler, D. R. and Towe, A. M., Genetics 47, 839-846, 1962.

Taylor, N. W. and Cluskey, J. E., Arch. Biochem. Biophys. 97, 399-405, 1962.

Wahba, A. J., Gardner, R. S., Basilio, C., Miller, R. S., Speyer, J. F. and Lengyel, P., Proc. Natl. Acad. Sci. 49, 116-122, 1963.

Weisblum, B., Benzer, S. and Holley, R. W., Proc. Natl. Acad. Sci. 48, 1449-1454, 1962.

Wollman, E. L. and Jacob, F., "La sexualité des bactéries", Masson & Cie, Paris, 1959, pp. 247.

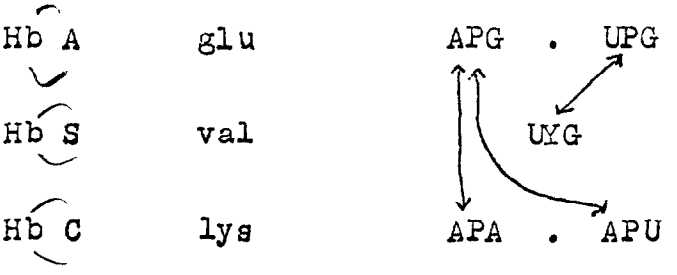
Zuckerkandl, E. and Pauling, L., in : "Horizons in Biochemistry", M. Kasha and B. Pullman, eds., Academic Press, New York, 1962, pp. 189-225.

Acknowledgment.

One of the authors (E.Z.) is greatly indebted to Professor Joshua Lederberg for discussions about the topics treated in this paper.

Fig. 1

The relation between coding triplets in hemoglobins A, S and C according to Eck's code.



Possible one-step transitions are marked by double arrows.
Middle letters of codons : P = purine, Y = pyrimidine.
Other symbols as usual (cf. Eck, 1963)